Data

1. Data Acquisition:

The data acquired for this project is a combination of data from three sources. The first data source of the project uses a London crime data that shows the crime per Borough in London. The dataset contains the following columns:

- lsoa_code: code for Lower Super Output Area in Greater London.
- Borough: Common name for London Borough.
- major_category: High level categorization of crime
- minor_category: Low level categorization of crime within major category.
- value: monthly reported count of categorical crime in given Borough
- year: Year of reported counts, 2008-2016
- month: Month of reported counts.

The second source of data is scraped from a Wikipedia page that contains the list of London Boroughs. This page contains additional information about the Boroughs with the following columns:

- Borough: The names of the 33 London Boroughs.
- Inner: Categorizing the Borough as an Inner London Borough or an Outer London Borough.
- Status: Categorizing the Borough as Royal, City or another Borough.
- Local authority: The local authority assigned to the Borough.
- Political control: The political party that control the Borough.
- Headquarters: Headquarters of the Boroughs.
- Area (sq mi): Area of the Borough in square miles.
- Population: The population in the Borough recorded during the year 2013.
- Co-ordinates: The latitude and longitude of the Boroughs.
- Nr. in map: The number assigned to each Borough to represent visually on a map.

The third data source is the list of Neighborhoods in the royal Borough of “Kingston upon Thames” as found on a Wikipedia page. This dataset is created from scratch using the list of neighborhoods available on the site with the following are columns:

- Neighborhood: Name of the neighborhood in the Borough.
- Borough: Name of the Borough.
- Latitude: Latitude of the Borough.
- Longitude: Longitude of the Borough.

2. Data Processing:

The data preparation for each of the three sources of data is done separately.
From the London crime data, the crimes during the most recent year (2016) are selected. The major categories of crime are pivoted to get the total crimes per Borough, per the category.

The second data is scraped from a Wikipedia page using the Beautiful Soup library in python. Using this library, we extract the data in the tabular format that is displayed on the website. After the web scraping, string manipulation is used to get the names of the Boroughs in the correct form. This is important because we will be merging the two datasets together on the column name “Borough”.

The two datasets are merged on the Borough names to form a new dataset that combines the necessary information in one dataset. The purpose of this dataset is to help in visualizing the crime rates in each Borough and identify the Borough with the least crimes recorded during the year 2016.

After visualizing the crime in each Borough, we can find the Borough with the lowest crime rate and hence tag that Borough as the safest Borough.

The third source of data is acquired from the list of neighborhoods in the safest Borough on Wikipedia. This dataset is created from scratch, the pandas data frame is created with the names of the neighborhoods and the name of the Borough. The ‘latitude’ and ‘longitude’ columns are left blank.

The coordinates of the neighborhoods are obtained using Google Maps Geocoding API to get the final dataset. The new dataset is used to generate the venues for each neighborhood using the Foursquare API.